Адаптация генеративных моделей к предметной области методом оптимизации низкоранговых матриц (LoRA) в реальном времени

Малафиевский Роман Сергеевич

Студент

МГУ имени М. В. Ломоносова, факультет ВМК, Москва, Россия

E-mail: s02220161@gse.cs.msu.ru

Научный руководитель *— Сучков Е. П.*

Многие задачи учебного и научного поиска формулируются как вопросы к предметной коллекции документов (учебники, статьи, конспекты). Современная практика построения систем вопрос-ответ для таких коллекций заключается в использовании Retrieval-Augmented Generation (RAG), где ответ генерируется языковой моделью на основе найденных фрагментов контекста. Однако даже при удачном извлечении контекста модель может слабо следовать терминологии источника, применять собственные, внешние знания, которых нет в источнике, и галлюцинировать.

В работе рассматривается подход *адаптации на этапе инференса* языковой модели на извлечённом тексте непосредственно перед генерацией ответа, без изменения изначальных весов. Подход ориентирован на контролируемое смещение добавленных весов так, чтобы повысить стилистическое и фактическое согласование ответов модели с источником.

LoRA (Low-Rank Adaptation) — метод адаптации модели посредством введения в отдельные линейные слои малоранговых добавок. Исходная матрица W не изменяется, вместо неё используется слагаемое низкого ранга ΔW:

$$\Delta W = \tfrac{\alpha}{r}\, BA \quad | \quad y = W'x = Wx + \tfrac{\alpha}{r}\, B\,(Ax) \quad | \quad A \in \mathbb{R}^{r \times d_{\text{in}}},\ B \in \mathbb{R}^{d_{\text{out}} \times r},$$

Постановка задачи. Рассматривается система вопрос-ответ над коллекцией документов $\mathcal{D}$. Для запроса q извлекается набор фрагментов $c = \text{Retrieve}(q, \mathcal{D})$, после чего модель генерирует ответ a. Сравниваются два режима: (1) *Base RAG*: генерация ответа на основе q и c без адаптации; (2) *TTT (Test-Time Training)*: генерация после короткой тестовой оптимизации LoRA-параметров на найденных фрагментах c, причём адаптер используется однократно и затем сбрасывается.

Ключевой исследовательский вопрос: даёт ли краткая LoRA-адаптация на извлечённом контексте измеримо и *управляемо* отличающееся поведение генерации по сравнению с базовым RAG, и как на это влияет выбор цели оптимизации и гиперпараметров.

Метод REEL (Retrieval-Enhanced Ephemeral LoRA). Предлагаемый алгоритм объединяет классический RAG (retrieval > reranking > generation) с кратковременной LoRA-адаптацией непосредственно перед генерацией. После извлечения текста к модели добавляется пустой LoRA-адаптер. Последовательно выполняется небольшое число шагов обучения на найденном тексте, затем генерируется ответ.

На шаге оптимизации обновляются только LoRA-параметры. Оптимизация проводится по задаче CLM (Causal Language Modeling) на информации из источника. Общий вид целевой функции:

$$\mathcal{L}(\varphi) \;=\; -\sum_{t=1}^{T} m_t \,\log p_{\theta,\varphi}\big(s_t \mid s_{<t}\big),$$

где θ — фиксированные параметры базовой модели, φ — LoRA-параметры, $m_t \in \{0,1\}$ — маска, задающая, какие токены участвуют в вычислении функции потерь (loss).

Шаг обновления LoRA-весов:

$$A_{k+1} = A_k - \eta\,\frac{\partial \mathcal{L}}{\partial A}, \qquad B_{k+1} = B_k - \eta\,\frac{\partial \mathcal{L}}{\partial B}, \qquad k = 0, \ldots, S-1.$$

Протокол и метрики оценки. Сравнивались три режима генерации: обычный RAG (без адаптации), REEL-Style (адаптация Causal LM по тексту извлеченных фрагментов) и REEL-Factual (префикс-запрос + контекст, где токены запроса маскируются при расчете loss). Оценка проводилась с помощью LLM-судьи (по критериям точности, фактологичности и стиля) и автоматических метрик привязки к источнику: ROUGE-L — близость ответа к источнику по самым длинным совпадающим последовательностям слов; Lexical overlap — доля содержательных слов (термины и т. п.) ответа, которые встречаются в исходном тексте; 4-gram — доля последовательностей из четырёх подряд идущих слов ответа, которые встречаются в исходном тексте; Supported rate — доля предложений ответа, для которых можно найти достаточно близкое по словам подтверждение в исходном тексте.

Результаты экспериментов. Эмпирическая оценка на корпусе

экзаменационных вопросов показала, что базовая модель (Qwen 4B) склонна опираться на параметрическую память: доля предложений, строго подкреплённых контекстом, составила лишь 29 %. Применение REEL обеспечило относительный прирост этой метрики на 22 % и стабильное улучшение ROUGE-L F1.

Практическая значимость. Продемонстрирован рабочий конвейер, объединяющий RAG с адаптацией на этапе инференса. Адаптация обратима: LoRA-параметры сбрасываются после каждого запроса, что исключает накопление ошибок. Метод не требует внешних обучающих данных и эффективно снижает долю фактически необоснованных утверждений (галлюцинаций), заставляя модель следовать терминологии и фактологии источника. Это делает подход перспективным для построения специализированных систем вопрос-ответ в узких предметных областях.

Шаги	**Режим**	**Авто-метрики (↑)**				**Оценка LLM-судьи (↑)**			
		Supp.	**ROUGE**	**Lex.**	**4-gr.**	**Fact.**	**Acc.**	**Style**	**Wins**
1	Обычный RAG	0.289	0.141	0.376	0.048	4.09	3.93	3.85	**18**
	REEL-Style	0.345	0.153	0.394	0.056	4.40	4.25	4.16	5
	REEL-Factual	0.345	0.152	0.393	0.055	4.36	4.24	4.07	2
3	Обычный RAG	0.289	0.141	0.376	0.048	4.09	3.93	3.85	**11**
	REEL-Style	0.346	**0.154**	0.397	0.059	4.42	4.27	**4.27**	7
	REEL-Factual	0.353	0.153	**0.400**	0.057	4.36	4.22	4.25	4
5	Обычный RAG	0.289	0.141	0.376	0.048	4.09	3.93	3.85	4
	REEL-Style	**0.366**	0.154	0.398	0.058	4.36	4.29	4.18	11
	REEL-Factual	0.344	0.154	0.397	**0.060**	**4.45**	**4.33**	4.25	**14**

Литература

1. Zhang T., Bi S., Hong Y., Zhang K., Luan F., Yang S., Sunkavalli K., Freeman W. T., Tan H. Test-time training done right // arXiv preprint arXiv:2505.23884. 2025.
2. Tandon A., Dalal K., Li X., Koceja D., Rød M., Buchanan S., Wang X., Leskovec J., Koyejo S., Hashimoto T., Guestrin C., McCaleb J., Choi Y., Sun Y. End-to-end test-time training for long context // arXiv preprint arXiv:2512.23675. 2025.